Optimization of Classification Accuracy Using K-Means and Genetic Algorithm by Integrating C4.5 Algorithm for Diagnosis Breast Cancer Disease

Fachrizal Ahdy Andoyo [1*], Riza Arifudin [1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: ahdyfachri@gmail.com

ARTICLE INFO

Article history
Received 8 February 2021
Revised 10 March 2021
Accepted 12 April 2021

Keywords
Data Mining
K-Means
Genetic Algorithm
Decision Tree
C4.5 Algorithm
Wisconsin Diagnostic Breast Cancer

ABSTRACT

Technological development resulted in data proliferating. The data is processed into valid information for daily needs. Data mining is a technique to convert data into useful information. Data mining has been widely used in performing prediction functions, for example, health and medical science. This study using Wisconsin Diagnostic Breast Cancer dataset taken from UCI Machine Learning Repository. The dataset has 32 attributes with 569 samples. This data has a continuous and high dimensional data type, and it makes the C4.5 algorithm need long computation time and extensive storage. This study aims to improve the accuracy of the C4.5 with a combination of K-Means and Genetic Algorithm. These study results compared the accuracy of the C4.5 algorithm before and after applying the combination of K-Means and the Genetic Algorithm for diagnosing breast cancer. The accuracy of C4.5 is 91,228%. Meanwhile, the accuracy of C4.5 after optimized using the K-Means and Genetic Algorithm is 94,824%, with the average number of features are selected 22 features. Thus, the application of K-Means and Genetic Algorithm on the C4.5 Algorithm can improve the accuracy of diagnosing breast cancer by 3,596%.

1 Introduction

The development of information and communication technologies makes humans completed their work more accessible (Sunge, 2018). However, this information is not necessarily accurate because the accurate information must use valid data and calculation processes. The data is processed into a subfield of science called data mining (Han, Kamber, & Pei, 2012). The application of data mining can be found in various fields, such as banking, marketing, insurance, urban planning, transportation, and medical and medical sciences (Sreedhar, Kasiviswanath, & Reddy, 2017). In medical science, data mining can be applied to diagnose breast cancer, diabetes, heart disease, *etc.* (Muslim *et al*., 2018). Breast cancer is one of the diseases that causes the highest mortality of women globally, characterized by the appearance of cells in the body growing out of control (Calle, 2005). Identification of breast cancer quickly and accurately is crucial for taking medical action. The process of identifying can use machine learning algorithms, such as classification techniques used to predict decisions in diagnosis, such as SVM, decision tree, KNN, Naïve Bayes, *etc.* (Huang, Li, & Ye, 2011).

The most widely used classification technique is the decision tree, a simple classification algorithm that is easy to use (Karegowda, Manjunath, & Jayaram, 2011). One of them is the C4.5 algorithm which can predict the best accuracy and minimum execution time (Muslim *et al*., 2018). However, the accuracy of the decision tree prediction can also be affected by continuous data

(Rajeshinigo & Jebamalar, 2017). The process needs a clustering technique to handle persistent data (Dubey, Gupta, & Jain, 2016).

Clustering is a process of grouping data into classes or groups so that objects in a group have high similarities compared to others (Karegowda *et al*., 2011). One of the clustering algorithms is the K-Means Algorithm. Class classifications involving high-dimensional data affect the computation of time and storage from the data processing stage that affect classification accuracy. A data dimension reduction method is used to handle high-dimension data, which is commonly called the selection feature (Talita, 2016).

The feature selection algorithm is part of the preprocessing data. The selection feature is also helpful for facilitating high-dimensional data processing (Wahyuni, 2016). Generally, the selection features are categorized into three categories, namely wrapped, embedded, and filter methods. The wrapped method requires ample computation time, memory space, and additional algorithms to produce the best subset (Talita, 2016). The filter method requires fast computation time to select features based on data characteristics, so it is not necessarily finding the best subset. The embedded form combines wrapped and filters to find the best combination of feature subsets (Boomert, Sun, & Bischl, 2020).

One of the wrapped algorithms that can optimize for accuracy is the Genetic Algorithm (Zamani, Amaliah, & Munif, 2012). Genetic algorithms are algorithms for solving solutions to problems based on the principles of natural selection in genetic science. When selecting features, the Genetic Algorithm is used as a random selection algorithm to explore large spaces (Arifudin, 2012). The purpose of the Genetic Algorithm is to choose the optimal value for weight by maintaining a population that has a good fitness value to produce offspring and form a new population (Alalayah, Almasani, & Qaid, 2018).

This study proposes combining the K-Means method and the Genetic Algorithm to optimize the classification C4.5 Algorithm in diagnosing breast cancer disease.

2 Methods

This study uses a combination K-Means and Genetic Algorithms as feature selection. K-Means is applied to solve continuous data that usually occurs in classification problems. Genetic algorithms are applied to find the best features based on the best fitness value at maximum generation. The combination of K-Means and Genetic Algorithm is used to improve classification accuracy. The classification method is the C4.5 algorithm. This study will determine the comparison of accuracy before and after the application of K-means and Genetic Algorithm in the C4.5 Algorithm. The flowchart of the proposed method is shown in Figure 1.

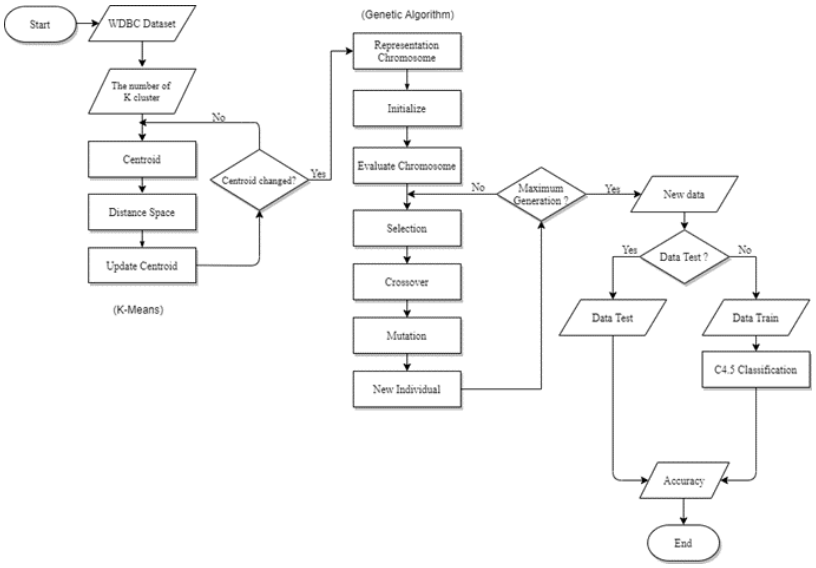

Figure 1. Flowchart C4.5 with K-Means and Genetic Algorithm

2.1 Data Preprocessing

This study using a public dataset, Wisconsin Diagnostic Breast Cancer, available in the UCI Machine learning repository. This dataset has ten attributes, where each feature has a criteria value calculated from each image. The attribute criteria are mean, standard error, and worst, thus make the overall attribute 30 and other attributes, namely ID and class. The dataset description can be seen in Table 1.

Table 1. Description of the WDBC dataset

No	Attribute	Description	Type
1	Radius (mean, se, worst)	The average distance from the center to the point on the perimeter	Continous
2	Texture (mean, se, worst)	Standard Deviation of the grayscale values	Continous
3	Perimeter (mean, se, worst)	Circumference of the cell nucleus	Continous
4	Area (mean, se, worst)	Cell nucleus area	Continous
5	Smoothness (mean, se, worst)	Local variation in long radius	Continous
6	Compactness (mean, se, worst)	Perimeter^2	Continous
7	Concavity (mean, se, worst)	The severity of the concave contour	Continous
8	Concave Point (mean, se, worst)	The number of concave sections of the contour	Continous
9	Symmetry (mean, se, worst)	Nuclear symmetry	Continous
10	Fractal Dimension (mean, se, worst)	Coastline approximation	Continous

2.1.1 K-Means

K-Means is a clustering algorithm. K-Means will divide the data into groups. This method is included in unsupervised learning, where the input received is data or objects and the number of k-clusters. The information is grouped based on the center point's value (centroid), representing the cluster or group.

K-means will classify existing data based on common characteristics. The k-means stages are as follows:

1. Determine the number of k-cluster.
2. Determine centroid randomly.
3. Calculate the distance between the data and the centroid using Equation 1.

$$D(x_2, x_1) = \sqrt{\sum_{i=1}^{n}(x_2 - x_1)^2} \quad (1)$$

Information:
D(x2, x1) : The data dimension
x1 : Position of the cluster center
x2 : Position of the data object

4. Group data based on data distance with the centroid.
5. Determine the value of the centroid using Equation 2.

$$c_k = \frac{1}{n_k}\sum d_i \quad (2)$$

6. Repeat steps 3-5 that the centroid values don't change anymore.

2.1.2 Genetic Algorithm

A genetic algorithm is an evolutionary method that solves problems using a random way. Inspired by natural selection, this method causes the variation to be collected in one direction, resulting in process optimization (Ashari, Muslim, & Alamsyah, 2016). The stages of the genetic algorithm in feature selection are as follows:

1. Initialize individuals to know what type of data will be used to calculate what will be represented on each chromosome.
2. Evaluate the fitness value of each particle in the population.
3. Selection, select chromosomes as prospective parents based on the fitness value of each chromosome. Chromosomes that have a good fitness value will be maintained.
4. Recombination produces a new chromosome with better fitness values than the previous chromosomes.
5. Mutation process, changes the value of genes on a chromosome.
6. Update the old chromosome value with the fitness value of the new chromosome.
7. Stop iteration if the best fitness value or maximum generation is met. Otherwise, go back to step 2.

2.2 C4.5 Algorithm

The C4.5 algorithm improves the IDE3 algorithm developed in 1986 by Quinlan Ross (Kathija, Nisha, & Sathik, 2017). In the C4.5 Algorithm, selection attribute using Gain, Ratio, by searching the entropy values (Sunge, 2018). C4.5 uses a decision model to determine the attributes that become the root by looking at the highest gain values of the existing attributes (Wibowo, Manongga, & Purnomo, 2020). The stages of the C4.5 algorithm in classifying the dataset are as follows:

1. Prepare training data that have been grouped into specific classes. Training data consist of 80% of the entire dataset.
2. Calculate the entropy value using Equation 3.

$$Entropy(S) = \sum_{i=1}^{n} -pi * log2pi \quad (3)$$

Information:
S : Case set
n : The number of S partitions
p_i : Proportion of S_i to S

3. Calculate Gain and splitEntropy using Equation 4 & 5

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} -\frac{|Si|}{|S|} * Entropy(Si) \quad (4)$$

$$SplitEntropy(S,A) = -\sum_{i+1}^{n} -\frac{|Si|}{|S|} * log2\frac{|Si|}{|S|} \quad (5)$$

Information:
S : Case set
A : Attribute
n : The number of A partition
$|S_i|$: The number of case on the i-partition
$|S|$: The number of case on the Case set

4. Calculate *GainRatio* using Equation 6

$$GainRatio(A) = \frac{Gain(A)}{SplitEntropy(A)} \quad (6)$$

5. Repeat step 2 until all records are partitioned
6. The decision tree partitioning will stop when:
    a. There are no attributes partitioned
    b. There is no record in the empty branch

3 Results and Discussion

This study uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, a public dataset from the UCI Machine Learning Repository, where this dataset has 32 attributes with 1 id, 1 class, and 30 attributes. This study uses algorithms that have been proposed and tested on the process. Test using the Python programming language and at the same time using libraries available in Python language. The results of this study are classification accuracy in diagnosing breast cancer.

3.1 Results

This study is divided into three applications. The first is to test the classification C4.5 algorithm. The C4.5 algorithm will process the WDBC dataset with 30 attributes based on the value of the gain ratio to produce accuracy in the percentage. The accuracy results from the classification C4.5 can be seen in Table 2.

Table 2. Result of C4.5 accuracy

Algorithm	Accuracy
C4.5 Algorithm	91,228 %

At the classification stage, the C4.5 algorithm produces an accuracy of 91,228 %. This process proves that the C4.5 can classify WDBC datasets with continuous data types well but can still be improved using preprocessing algorithms.

The second application is a combination of algorithm C4.5 with K-Means to process the dataset WDBC. K-Means will create groups or clusters on attributes that have continuous data. In the K-Means process, determining the number of k-clusters will affect accuracy. The results of the accuracy of the model combinations can be seen in Table 3.

Table 3. Result of C4.5 accuracy with K-Means

The number of k	Accuracy
2	93,859%
3	89,473%
4	92,982%
5	91,228%

The third application is the C4.5 algorithms will be combined with K-Means and Genetic Algorithms. The genetic algorithm will search for the best features based on the best fitness values in this process. This test was carried out ten times. The application of this combination will result in the best parts to be selected and the classification accuracy. The results of the accuracy of the model combinations can be seen in Table 4.

Table 4. Average C4.5 Accuracy Results with a combination of K-Means and Genetic Algorithms

Execution	Accuracy	Feature selected
1	95,614%	17
2	94,736%	30
3	90,350%	23
4	95,614%	19
5	93,859%	20
6	95,614%	23
7	94,736%	20
8	95,614%	28
9	96,491%	18
10	95,614%	24
Average	**94,824%**	**22,2**

3.2 Discussion

This study's purpose is to determine the process and result of the accuracy of combination K-Means and Genetic Algorithm on C4.5 algorithm for diagnosis breast cancer disease. K-Means are applied to attributes that have a continuous data type. The choice of k-cluster will affect the accuracy. Based on experiments conducted with k=2,3,4, and 5. In this case, k=2 has the best accuracy in optimizing the C4.5 algorithm.

The following process is selecting the best feature using Genetic Algorithm based on the best fitness values. The accuracy obtained from this combination is 94,824%. So these results state that this combination can increase the accuracy of the C4.5 algorithm by 3,596%. The comparison of the accuracy results before and after the application algorithm can be seen in Figure 2.

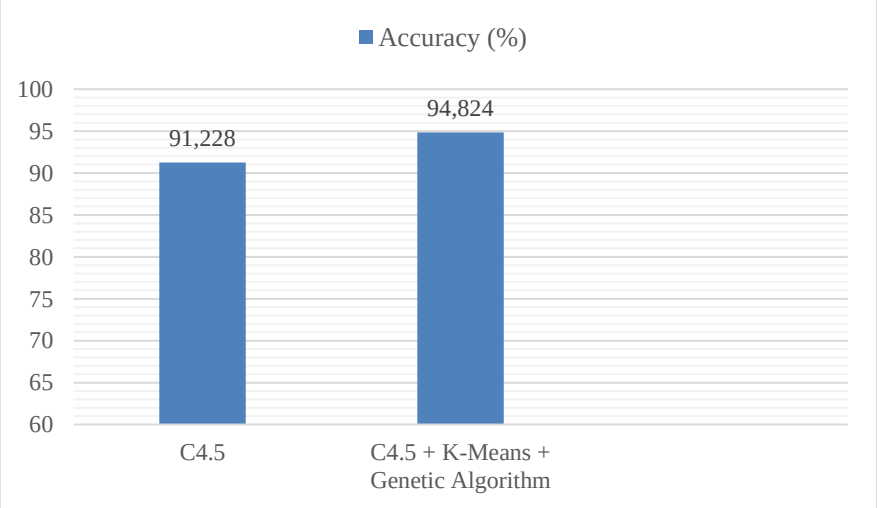

Figure 2. Comparison of the accuracy

Based on the study result from the application of a combination of K-Means and Genetic Algorithm to optimize the accuracy of the C4.5 algorithm for diagnosis of breast cancer disease that has been done, it can be seen that the comparison of the accuracy before and after C4.5 is optimized

using a combination of K-Means and Genetic Algorithm. In contrast, the comparison of this study with the previous research can be seen in Table 5.

Table 5. The related research

Researchers	Year	Methods	Result
Hermawanti	2012	C4.5	94,56%
Zamani *et al.*,	2012	GA+Neural Network	97%
Elouedi *et al.*,	2014	K-Means+C4.5	K-Means 91,56%
			K-Means+C4.5 95,14%
Rajeshinigo *et al.*,	2015	K-Means+C4.5	92%
Muslim *et al.*,	2018	PSO+C4.5	C4.5 95,61%
			PSO+C4.5 96,49%

The advantages of the study through this model by applying a combination of the K-Means and Genetic Algorithm as selection feature in the classification of the C4.5 algorithm can increase the accuracy for diagnosing breast cancer disease. While the study's weakness is that the accuracy result tends to change because it depends on the initialization for determining the fitness values of the Genetic Algorithm is selected randomly.

4 Conclusion

In this study, the application of the C4.5 algorithm combines with K-Means and Genetic Algorithm as a feature selection for the diagnosis WDBC dataset obtained from the UCI Machine Learning Repository. K-Means implemented to handle the problem of continuous data on the attributes of the WDBC dataset. At the same time, the Genetics Algorithm is applied to choose the best features based on the best fitness value. This study resulted in the accuracy of the application of the algorithm C4.5 is 91,228%. The preprocessing algorithm that can improve the results of the C4.5 algorithm is to use a combination of K-Means where the best k-cluster is k = 2 and combine with Genetics Algorithm as feature selection resulting in an accuracy of 94.824%. Thus, this study can be concluded that applying the K-Means and Genetics Algorithm at C4.5 algorithm can improve results accuracy in diagnosing breast cancer by 3,596 %.

References

Alalayah, K. M. A., Almasani, S. A. M., & Qaid, W. A. A. (2018). Breast Cancer Diagnosis Based on Genetic Algorithms and Neural Networks. *International Journal of Computer Applications*, 180(26), 42-44. doi: 10.5120/ijca2018916605

Arifudin, R. (2012). Optimasi Penjadwalan Proyek dengan Penyeimbangan Biaya Menggunakan Kombinasi CPM dan Algoritma Genetika [Optimization of Project Scheduling with Cost Balancing Using a Combination of CPM and Genetic Algorithms]. *Jurnal Masyarakat Informatika*, 2(4), 1-14. doi: 10.14710/jmasif.2.4.1-14

Ashari, I. A., Muslim, M. A., & Alamsyah. (2016). Comparison Performance of Genetic Algorithm and Ant Colony Optimization in Course Scheduling Optimizing. *Scientific Journal of Informatics*, 3(2), 149-158. doi: 10.15294/sji.v3i2.7911

Boomert, A., Sun, X., & Bischl, B. (2020). Benchmark for Filter Methods for Feature Selection in High-Dimensional Classification Data. *Computational Statistics and Data Analysis*, 143, 1-19. doi: 10.1016/j.csda.2019.106839

Calle, J. (2005). *Breast cancer facts and figures 2005-2006*. Atlanta: American Cancer Society.

Dubey, A. K., Gupta, U., & Jain, S. (2016). Analysis of K-means Clustering Approach on the Breast Cancer Wisconsin Dataset. *International Journal of Computer Assisted Radiology and Surgery*, 11(11), 2033-2048. doi: 10.1007/s11548-016-1437-9

Han, J., Kamber, M, & Pei, J. (2012). *Data Mining: Concepts and Techniques (3nd ed).* Waltham: Morgan Kaufmann.

Huang, Y., Li, W., & Ye, X. (2011). A study of Genetic Neural Network as Classifiers and its Application in Breast Cancer Diagnosis. *Journal of Computers*, 6(7). 1319-1324. doi: 10.4304/jcp.6.7.1319-1324

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2011). Application of Genetic Algorithm Optimized Neural Network Connections Weights for Medical Diagnosis of Pima Indians Diabetes. *International Journal on Soft Computing (IJSC)*. 2(2). 15-23. doi: 10.5121/ijsc.2011.2202

Kathija, A., Nisha, S. S., & Sathik, M. M. (2017). Classification of Breast Cancer Data Using C4.5 Classifier Algorithm. *International Journal of Recent Engineering Research and Development (IJRERD)*, 2(2), 13-19.

Muslim, M. A., Rukmana, S. H., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2018). Optimization of C4.5 Algorithm-based Particle Swarm Optimization for Breast Cancer Diagnosis. *Journal of Physics: Conferences Series,* 983, 1-7. doi: 10.1088/1742-6596/983/1/012063

Rajeshinigo, D., & Jebamalar, J. P. A. (2017). Accuracy Improvement of C4.5 using K-means Clustering. *International Journal of Science and Research (IJSR)*, 6(6), 2755-2758.

Sreedhar, C., Kasiviswanath, N., & Reddy, P. C. (2017). Clustering Large Datasets using K-means Modified Inter and Intra Clustering (KM-I2C) in Hadoop. *Journal of Big Data,* 4(27), 1-19. doi: 10.1186/s40537-017-0087-2

Sunge, A. S. (2018). Optimasi Algoritma C4.5 Dalam Prediksi Web Phishing Menggunakan Seleksi Fitur Genetic Algoritma [Optimization of C4.5 Algorithm in Predicting Web Phishing Using Genetic Algorithm Feature Selection]. *Paradigma*, 20(2), 27-32. doi: 10.31294/p.v20i2.4021

Talita, A. S. (2016). Klasifikasi Wisconsin Diagnostic Breast Cancer Data dengan Menggunakan Sequential Feature Selection dan Possibilistic C-Means [Wisconsis Diagnostic Breast Cancer Data Classification using Sequential Feature Selection and C-Means Possibilitic]. *Jurnal Ilmiah Komputasi*. 15(1). 47-52. doi: 10.32409/jikstik.15.1.144

Wahyuni, E. S. (2016). Penerapan Metode Seleksi Fitur untuk Meningkatkan Hasil Diagnosis Kanker Payudara [Feature Selection Method Implementation to Increase Breast Cancer Diagnostic Results]. *Jurnal Simetris*, 7(1), 283-294. doi: 10.24176/simet.v7i1.516

Wibowo, A., Manongga, D., & Purnomo, H.D. (2020). The Utilization of Naïve Bayes and C4.5 in Predicting the Timeliness of Student's Graduation. *Scientific Journal of Informatics*, 7(1), 99-112. doi: 10.15294/sji.v7i1.24241

Zamani, A. M., Amaliah, B., & Munif, A. (2012). Implementasi Algoritma Genetika pada Struktur Backpropagation Neural Network untuk Klasifikasi Kanker Payudara [Genetic Algorithm Implementation in Backpropagration Neural Network Structure for Breast Cancer Classification]. *Jurnal Teknik ITS*, (1), 222-227. doi: 10.12962/j23373539.v1i1.638